





Deep survival modeling of longitudinal retinal OCT volumes for predicting the onset of atrophy in patients with intermediate AMD

ANTOINE RIVAIL,¹  WOLF-DIETER VOGL,² SOPHIE RIEDL,²
CHRISTOPH GRECHENIG,² LEONARD M. COULIBALY,² GREGOR S.
REITER,² ROBYN H. GUYMER,^{3,4} ZHICHAO WU,^{3,4} URSULA
SCHMIDT-ERFURTH,² AND HRVOJE BOGUNOVIĆ^{1,*} 

¹Christian Doppler Lab for Artificial Intelligence in Retina, Department of Ophthalmology and Optometry, Medical University of Vienna, Vienna, Austria

²Department of Ophthalmology and Optometry, Medical University of Vienna, Vienna, Austria

³Centre for Eye Research Australia, Royal Victorian Eye and Ear Hospital, East Melbourne, Australia

⁴Department of Surgery (Ophthalmology), The University of Melbourne, Melbourne, Australia

*hrvoje.bogunovic@meduniwien.ac.at

Abstract: In patients with age-related macular degeneration (AMD), the risk of progression to late stages is highly heterogeneous, and the prognostic imaging biomarkers remain unclear. We propose a deep survival model to predict the progression towards the late atrophic stage of AMD. The model combines the advantages of survival modelling, accounting for time-to-event and censoring, and the advantages of deep learning, generating prediction from raw 3D OCT scans, without the need for extracting a predefined set of quantitative biomarkers. We demonstrate, in an extensive set of evaluations, based on two large longitudinal datasets with 231 eyes from 121 patients for internal evaluation, and 280 eyes from 140 patients for the external evaluation, that this model improves the risk estimation performance over standard deep learning classification models.

© 2023 Optica Publishing Group under the terms of the [Optica Open Access Publishing Agreement](#)

1. Introduction

Age-related macular degeneration (AMD) is a leading cause of severe, irreversible vision loss in people over age 60, with an estimated 196 million patients affected worldwide in 2020 [1]. AMD slowly progresses from early/intermediate to late stage of the disease, where central vision loss can occur. Late AMD includes neovascularization or atrophy of the retina, with different underlying pathomechanisms. Although the early and intermediate stages of the disease typically carry no visual symptoms, morphological changes in the outer retina can be observed. In particular, the atrophic stage is characterized by gradual morphological deterioration of the outer retina and the thinning and loss of the photoreceptors [2].

Color fundus photography (CFP) and optical coherence tomography (OCT) are the most common imaging modalities used to detect and monitor AMD. Especially OCT, being a 3D imaging modality with a micrometer-scale resolution, allows studying the morphological changes associated with the disease in great detail. Nevertheless, in patients with intermediate AMD, the risk of progression to late stages is highly heterogeneous, and the prognostic imaging biomarkers remain unclear. There is therefore an unmet need to develop predictive models for estimating the risk of the onset of late AMD. Having an accurate risk estimators would also help improve the understanding of the underlying pathomorphological mechanisms.

The development of predictive models of AMD relies on the availability of longitudinal imaging data. The most common source of such data, eyes with drusen and no late AMD, comes from observing the fellow eyes of patients treated for neovascular AMD [3,4], from the control

arms of intermediate AMD intervention studies [5], or from the observational longitudinal studies [6,7]. However, the slow nature of AMD progression requires very long observation times, and in the studies with only a few years of duration, only a small minority of patients are observed progressing to late AMD.

To tackle the above problems and obtain effective risk estimation predictive models, we combine two powerful techniques: survival analysis and deep learning. Survival analysis is specifically designed to handle longitudinal data, as it is able to model both the occurrence of events and their censoring when the event time is unknown. However, traditional survival models are often limited in terms of modeling complexity, as they require a representation in the form of a predefined set of quantitative imaging biomarkers, which are often difficult to extract or are unknown a priori.

On the other hand, the developments in the field of artificial intelligence (AI) and especially deep learning, provided the capability to learn effective representations directly from the raw imaging data. Yet, the majority of deep learning in retina focuses on prediction from 2D images rather than 3D volumes, and on classification or regression tasks, rather than risk estimation or time-to-event prediction.

1.1. *Related work*

Predictive modeling is a key task in medical imaging and has been applied to many clinical problems in the field of ophthalmology, as it is essential to identify patients at risk, as well as identify novel risk factors. The task is often treated as an image classification problem. In [8], qualitative and quantitative OCT features were generated at baseline to train multivariate logistic regression models to estimate the risk of progression from intermediate to non-neovascular atrophic AMD. In [9], a large set of quantitative imaging biomarkers were automatically extracted and used as covariates to predict the progression to a late AMD stage with a generalized linear model. Similarly, in [7], qualitative features were used to provide risk estimates using decision trees. These predictive models suffer from low model capacity and cannot handle high-dimensional data, therefore they are limited to a set of quantitative features and are not able to exploit new features in the form of imaging patterns present in raw image data. In contrast, deep learning allows building models directly from raw image data and overcome these limitations.

Several deep learning models have been developed for predicting progression from intermediate to late AMD. For instance, Russakoff et al. [10] developed a specific deep learning network for 2D B-scan classification, called AMDNet, to predict whether eyes are likely to progress to wet AMD. Banerjee [11] proposed a deep recurrent network for predicting GA progression from a time series of a set of 21 OCT imaging biomarkers and evaluated it for multiple prediction intervals. A two-stage system for prediction of progression to neovascular AMD from OCT was explored in [4], which consists of two serially-connected neural networks. The first one segments the clinically important features in OCT, and the segmentation maps are provided to the second, classification network. Such an approach simplifies the model interpretation and makes the second stage device-agnostic. Bora et al. [12] described a deep neural network classifier that predicts a risk of developing diabetic retinopathy from color fundus images. However, the above methods operate in a classification setting, which ignores the temporal information in the data, namely the time-to-event (progression time point or censoring) is not taken into account.

Survival models are an essential tool for modeling longitudinal medical data. They allow estimating risks, and correctly accounting for censoring. A common family of survival models are based on the linear Cox Proportional Hazard (CoxPH) model [13], and have previously been applied to AMD progression prediction from a small set of clinically relevant quantitative imaging biomarkers [14]. Adaptation have been proposed to extend survival models to non-linear models, such as piece-wise exponential models [15] or early survival neural networks [16,17]. Another approach, denoted as Random Survival Forest, is an adaptation of the random forest with

a specific survival-based decision tree branching rule that allows for the analysis of right-censored survival data [18].

Recently, survival methods have been adapted to standard deep learning architectures. In [19], the authors combine a deep convolutional neural network (CNN) feature extractor with a traditional linear CoxPH model. Similarly, DeepSurv [20] and DeepConvSurv [21] are based on the Cox model where the linear part is replaced by a fully connected network or a CNN, respectively. Those models share the same limitations as the standard Cox models, namely the proportionality and time independence assumptions. On the other hand, DeepHit [22] and Logistic Hazard (LH) [23] model directly the survival times without the previous assumptions and allow competing risks.

1.2. Contribution

In this paper, we explore combining the strengths of survival statistics and deep learning in the form of a deep survival prediction algorithm for retinal OCT. This method is applied to the problem of predicting progression from intermediate to late atrophic AMD from longitudinal 3D retinal volumes (Fig. 1).

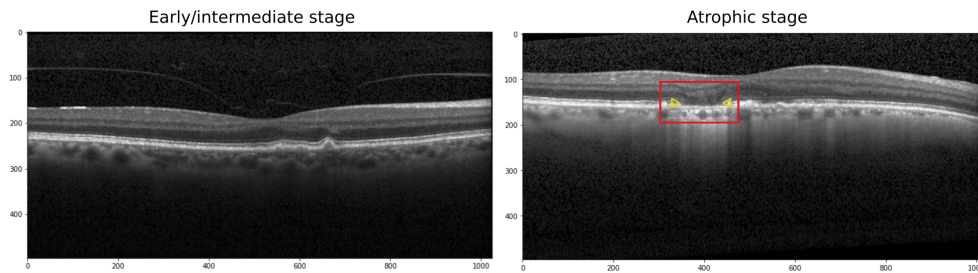


Fig. 1. Example of a retinal OCT showing conversion from early/intermediate AMD (left) to the first OCT signs of atrophy (right). The atrophic lesion (red box) is where the loss of outer retinal layers (subsidence of the outer plexiform and inner nuclear layer and hyperreflective wedge shaped bands in Henle's nerve fiber layer), loss of the retinal pigment epithelium and hypertransmission of the signal into the choroid are noticeable. The yellow triangles denote the extremities of the atrophic lesion.

The proposed method allows to correctly account for the time-to-event and the right-censoring of patients, while learning from raw 3D image data, and adapts the LH loss presented in [23] to longitudinal retinal OCT. In brief, the contributions of this paper are the following:

1. We compare and quantitatively evaluate the benefit of using a deep learning survival loss in contrast to a standard binary cross entropy loss for prediction from raw OCT data, and to a traditional Cox proportional hazards (CoxPH) model trained from a set of quantitative imaging biomarkers.
2. The predictive model operates on 3D OCT volumes, instead of 2D images, and the performance and the generalization of the model is assessed on two large longitudinal datasets of patients with intermediate AMD.
3. We analyze the best-performing deep learning model with post-hoc interpretability methods to investigate the predictive role of different retinal regions for a progression to atrophy.

2. Methods

2.1. Predictive model formulation

In a longitudinal dataset, for each patient eye, a discrete set of observations are available for the visit times $t_0 \cdots t_N$. After the last observation, the eye is considered as positive if a progression from intermediate to atrophic AMD stage occurred at a particular visit, or as censored if the eye remained in intermediate AMD stage within the study interval. We address the task of predicting the progression from retinal OCT scans with two different deep learning-based methodological approaches, where we consider it either as a survival modelling task or as a binary classification task (Fig. 2).

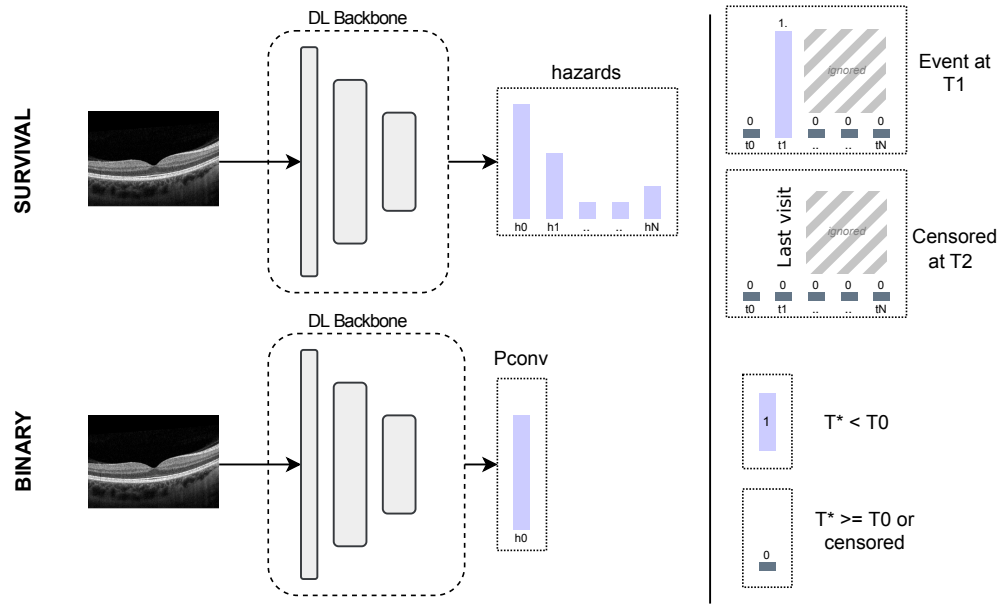


Fig. 2. The difference between survival and binary classification setting, in terms of labels and predicted output. In the survival setting with logistic hazard (LH) loss, the time-to-event is encoded in the target label with a single value of one at the event time, or all zero values for a censored patient. In the binary classification setting, a threshold for the time-to-event is defined to obtain a binary label, a positive one for a patient progressing before T_0 , and a negative one otherwise.

Survival modeling task: Let T^* be the time point when the event was observed and ΔT the interval between two visits. The occurrence of this event is described with the following functions in a survival setting: the probability mass function (PMF), f , which describes the probability that the event occurs at a specific time point (between t_j and $t_j + \Delta T$); the survival function, S , which is the probability that the event occurs after a given time; and the hazard rate, h , as the probability that the event occurs between t_{j-1} and t_j . They are defined in the following way:

$$\begin{aligned}
 f(t_j) &= P(T^* = t_j) \\
 S(t_j) &= P(T^* > t_j) = \sum_{k>j} f(t_k) \\
 h(t_j) &= P(T^* = t_j | T^* > t_{j-1}) = \frac{f(t_j)}{S(t_{j-1})}.
 \end{aligned} \tag{1}$$

Traditional or deep learning-based survival models allow estimating the hazards or survival probabilities based on the available observations.

Binary classification task: Let T_0 be the time interval of clinical relevance. The observation is considered as a positive case when the progression happens before T_0 , and a negative otherwise. Deep learning models can then be trained for this task using a standard binary cross-entropy loss. We want to remark that the binarization of the time-to-event results in a partial loss of information. Indeed, all progressors regardless of the time to progression are in the same category, similarly for censored patients, where the time to censoring is ignored. Thus, the task consists of determining the probability $P_{\text{progression}}$ of progressing to a late atrophic AMD stage within the predefined interval:

$$P_{\text{progression}} = P(T^* \leq T_0) \quad (2)$$

It is worth noting that the probability of progression can be directly linked with the survival probability:

$$\begin{aligned} P_{\text{progression}} &= P(T^* \leq T_0) \\ P(T^* \leq T_0) &= 1 - P(T^* > T_0) = 1 - S(T_0) \\ P_{\text{progression}} &= 1 - S(T_0) \end{aligned} \quad (3)$$

It shows explicitly that the survival model tackles all possible intervals at once. We can also switch from survival to binary setting for the evaluations in our experiments.

2.2. Survival modeling adaptation for deep learning

The deep learning backbone in the form of a CNN allows processing raw 3D OCT data, and to learn through training to detect the imaging patterns relevant for the prediction tasks. However, to build a deep learning survival model, we need an adapted training loss. Our work is based on the *logistic hazard* (LH) loss [23], chosen because it does not require assumptions on the hazard distribution. It is defined as:

$$L = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{k(t_i)} [y_{ij} \log(h(t_j|x_i)) + (1 - y_{ij}) \log(1 - h(t_j|x_i))], \quad (4)$$

where x_i is the available observation for patient i , $h(t_j|x_i)$ is the predicted hazard for time t_j given the observation x_i , y_{ij} indicates whether an event is observed at time t_j for patient i , and $k(t_i)$ is the index of the first event (progression or censoring). This loss represents the negative log-likelihood of the observed events in terms of hazards. The hazards after the event ($j > k(t_i)$) are ignored to account for censoring. The final loss resembles the binary cross-entropy, except the hazards are parametrized with a neural network using a sigmoid activation function and an additional specific censoring mask. Each label has N_{visit} variables (in our case 6 times 12-month intervals), and it contains a single value 1 at the time of the event, and all zero values if censored (Fig. 3).

2.3. Volume preprocessing

An illustration of the applied OCT preprocessing procedure is shown in Fig. 4. All 3D OCT volumes were flattened according to the inner limiting membrane (ILM) layer, obtained with automated segmentation by IOWA reference algorithms [24,25]. ILM was chosen because it is a robust segmentation target, and any resulting segmentation errors are distant from the primary region of interest, which in the case of AMD is in the outer retina. The 3D OCT volumes were then cropped to a fixed physical size (3.8 mm in depth along B-scans, 5.6 mm in width, 0.62 mm in height along A-scans) centered on the fovea and resized to $32 \times 512 \times 160$ voxels, respectively.

The height dimension was cropped to keep only the clinically relevant region composed of retinal layers (ILM to Bruch's membrane) and the choroid. Similarly, the peripheral B-scans were removed, as the atrophy of the retina often starts in the central region. The cropping and

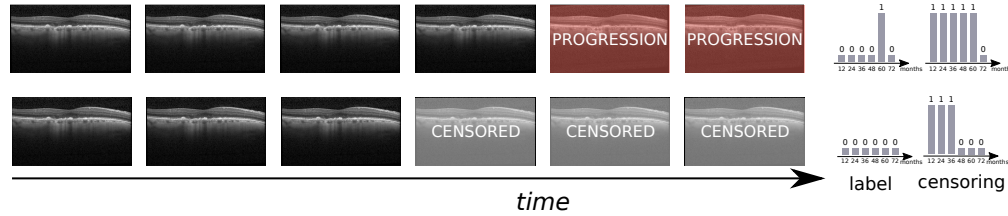


Fig. 3. Example of time-to-event labeling. Labels for hazards and censoring are used for the training. In the upper part of the figure, examples of label and censoring vectors are displayed. The label indicates the time of progression as a one-hot vector and the censoring vector the available visits before censoring.

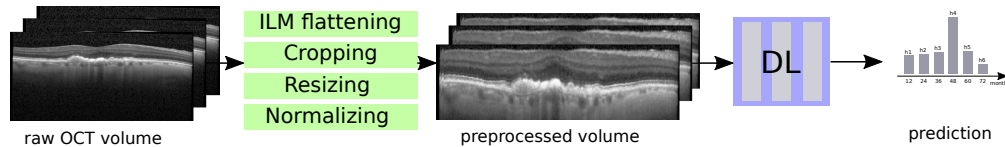


Fig. 4. Example of OCT volume preprocessing steps before supplying the volume to the deep learning (DL) model. The OCT volume is flattened using ILM layer segmentation, and then cropped to a common field of view, resized, and intensity normalized.

resizing was done mainly to reduce the memory (VRAM) footprint required for the training on a GPU. The intensity values were normalized to have a zero mean and unit standard deviation per OCT volume.

3. Experimental setup

3.1. Datasets

Participants For our experiments, we used data from two longitudinal imaging studies of AMD patients with eyes having drusen and no late AMD at baseline (Table 1). The first dataset, *MUV*, consists of OCT scans of patients with at least one eye with drusen and no late AMD, part of a long-term observational study at the Department of Ophthalmology, Medical University of Vienna (MedUni Wien) [6]. It is composed of 231 eyes from 121 patients (91% with early/intermediate AMD, i.e., bilateral drusen). Each eye was imaged with a 3- to 6-month intervals, and a follow-up duration ranged from two to seven years. The second dataset is obtained from the sham arm of the *LEAD* study [5], and includes well curated OCTs of patients with bilateral large drusen AMD (a subset of iAMD) at baseline. The LEAD dataset consists of 280 eyes from 140 patients, and each eye was imaged with 6-month intervals for a study duration of 3 years. The cumulative number of converters over time in both datasets is displayed in Fig. 5.

Table 1. Properties of the employed longitudinal OCT datasets. It describes the number of patients, eyes, OCT scans, and number of eyes that progressed, as well as the duration of the follow-up and the interval between visits.

Dataset	Eyes/Patients	OCTs	progressors	Visit interval	Visits	Duration
MUV	231/121	3066	38	3-6 months	3+	53 ± 34 months
LEAD	280/140	1531	47	6 months	7	36 months

All patients gave informed consent prior to inclusion in the respective studies. This retrospective analysis was approved by the Ethics Committee at MedUni Wien (EK Nr: 1246/2016). All study

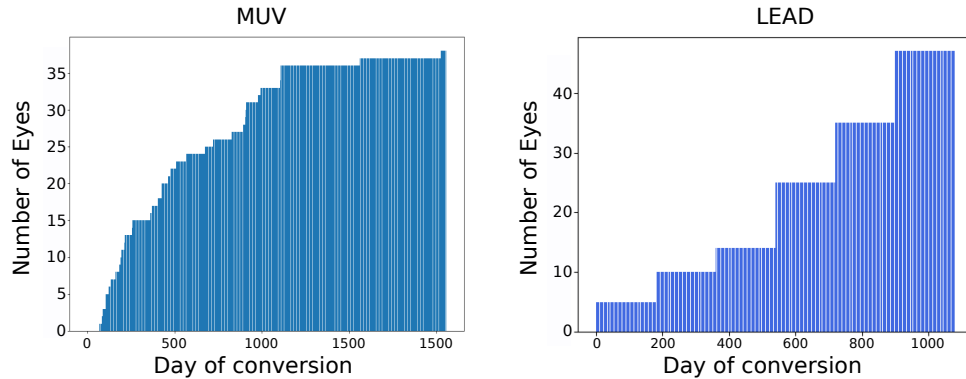


Fig. 5. Cumulative number of converted eyes with respect to the time to progression in the MUV dataset (left) and LEAD dataset (right).

procedures were conducted in accordance with the Declaration of Helsinki, and all the patient data were pseudonymized.

OCT imaging and grading of AMD progression Both datasets consisted of OCTs scans acquired with a Spectralis scanner (Heidelberg Engineering, Heidelberg, DE). The OCT volumes consisted of 49 B-scans having $512 \times 1024 \times 496$ pixels, and covered an en-face field of view of $20^\circ \times 20^\circ$. Progression to late atrophic AMD was considered as soon as complete RPE and outer retinal atrophy (cRORA) was detectable on OCT. We adopted the cRORA definition given by the classification of atrophy meetings (CAM) program [2]. Therefore, the following features needed to be present: at least $250 \mu\text{m}$ zone of hypertransmission (1) and disruption of RPE of at least $250 \mu\text{m}$ (2) and an additional evidence of overlying photoreceptor degeneration (3). To identify the time point of progression to late AMD, the OCT scans of both datasets were analyzed and (re)graded by the same team of retinal experts at MedUni Wien for consistency. Eyes that had atrophy but did not fulfil the cRORA criteria, as well as those that developed only neovascular AMD, were considered not to have progressed.

The MUV dataset, where a progression to late atrophic AMD was observed in 41 eyes, was used for training and internal evaluation of the predictive models. It was divided into five folds to perform cross-validation, with three folds for training, one for validation, and one for testing. The split was performed at the eye level, and the folds were stratified with respect to the proportion of progressors to ensure their representativeness, given the small number of progressing eyes. In contrast, LEAD dataset, where 47 eyes progressed to late atrophic AMD, was used exclusively as an external validation set.

3.2. Deep learning setup

All deep learning experiments are based on the 3D ResNet18 CNN backbone as defined in [26], pretrained on Kinetics-400 video action recognition dataset. Performing the training and the prediction in 3D allows to directly use the available volume-level progression labels.

The chosen 3D pretrained ResNet18 gave a good trade-off between capacity and trainability given the size of our datasets. Because of the low ratio of converters in MUV and LEAD dataset with around 16% of progressors, the censored patients were randomly under-sampled by 50% in the training set to counteract the imbalance. We kept the original ratio for the validation and the test sets to get a performance estimation representative of the study cohort. During training, the following data augmentation was applied: random cropping, flip (vertical axis), random rotation and contrast augmentation. The model was trained with AdamW gradient descent

method [27]. The optimization parameters (learning rate, β_1 , β_2 and weight decay) were chosen through hyperparameter tuning using a Python library Ray Tune. The model selection was based on picking the best epoch score on the validation set.

3.3. Evaluation procedure

Method baselines: We compare our proposed deep survival model (Deep LH) with two other baselines.

1. Deep binary classification models, which are trained to predict a progression within a specific time interval. They output a single value, corresponding to the probability of a progression occurring within the target interval. We trained the model for 12, 24 and 36-month intervals. Models are trained with binary cross-entropy and are denoted as Deep BCE models.
2. CoxPH model [13] based on a set of quantitative OCT imaging and demographic features. These imaging features were extracted from the OCT images using a set of deep learning-based image segmentation algorithms. This followed the approach of [14] but excluding the manual grading of color fundus images. The input vector contains seven features: baseline age, three drusen related features (volume, height variability, and mean reflectivity) and three features of hyperreflective foci volume (in three separate retinal layers). The model was trained using the Python library scikit-survival [28] (version 0.12.0).

To provide a fair comparison, all models were trained on the same cross-validation folds. To evaluate our models, we relied on two relevant metrics for prediction problems:

Dynamic AUC: Dynamic or cumulative area under receiver operating characteristic (AUC) allows extending the standard AUC for time-dependent measures [29]. The dynamic AUC consists of computing the AUC at different time intervals, in our experiments, 12, 24 and 36 months. The average AUC is then calculated as a measure of the performance across the three time intervals.

For classification models, the output probability of progression was used as a general risk score. For survival models, which output a time-dependent survival prediction, the dynamic AUC was evaluated for the corresponding survival probability at that time.

Concordance Index: To evaluate the predicted survival function, we used the concordance index (CCI), which accounts for censored data and describes the model's ability to rank patients given their actual risk. Concordance index allows to evaluate individual risk scores either from the predicted hazards (Deep LH model) or from the binary probability (Deep BCE models). For Deep LH models, the individual risk score is taken as the progression probability at 3 years, which was selected based on predictions on the validation set. For Deep BCE models, we also compute the CCI, by using the progression probability as a risk score.

Statistical analysis: To compare the CCI and dynamic AUC scores from different models, the confidence intervals were obtained by bootstrapping the predictions with 1000 resamples at the patient level, and Wilcoxon signed-rank test was applied to test differences. For the external dataset, a single prediction is obtained by ensembling the 5 models resulting from the cross-validation on the internal dataset. To compare survival curves in the risk groups experiment, we used the logrank test with the `statsmodel` library [30].

3.4. List of experiments

Internal evaluation: Cross-validation was performed on the MUV dataset, for all models: proposed Deep LH model, and baselines: Deep BCE models, and CoxPH model. The predictions were evaluated with both dynamic AUC and CCI. Dynamic AUC allows evaluating the models in a classification setting for different intervals. On the other hand, CCI evaluates the predicted

risks of patients given their actual progression or censoring time. For each eye, all visits were used for validation and test set metrics.

External validation: To further evaluate the generalization capabilities, the performance of Deep LH, Deep BCE models and the CoxPH models were evaluated with dynamic AUC and CCI on the external validation dataset LEAD. For each sample, the predictions of the five models from the cross-validation on MUV dataset were averaged to get a single prediction score. Analogous to internal validation, all visits were considered for performance evaluation.

Risk group analysis: To verify the stability of the predicted survival probabilities, the thresholds defining four risk groups were generated based on the risk quartiles on the validation set. The eyes in the test set were split into the four risk groups using these thresholds, and the corresponding Kaplan-Meier curve estimates were computed. Statistical difference between low risk and high risk group was tested.

Model interpretability: To understand the imaging patterns Deep LH model relies on for its prediction, we first conducted an occlusion analysis. There, specific image regions were first masked out in the OCT volumes (Fig. 8). Then, we computed the cross-validation performance and compared it to the original dataset, where the relative drop in performance reveals the extent to which each region contributes evidence to the prediction. The following regions were masked: the *fovea* (central 1 mm region), the *extrafovea* (outside central 1 mm region), the *choroid* (below the Bruch's membrane), the *inner* retina (between ILM and INL-OPL) and the *outer* retina (between INL-OPL and the Bruch's membrane). In addition, because the occlusions may introduce artifacts around their boundaries, we also computed Class Activation Maps (CAM) [31]. In the case of Deep LH model, we obtain a CAM image for each hazard bin, i.e., every 12 months.

4. Results

4.1. Internal evaluation

The results of the internal evaluation are presented in Table 2 and the dynamic AUC curves are plotted in Fig. 6. The CoxPH obtained slightly better performance in terms of average dynamic AUC with a mean value of 0.80 (95 % CI [0.749 - 0.849]) and CCI with a value of 0.78 (95 % CI [0.735 - 0.833]). However, Deep LH models obtained similar performance but outperformed clearly all Deep BCE models. The latter produced mixed results, with the highest performance achieved for 36-month training interval. Such training with a fixed prediction interval showed its limits on this dataset with large amount of censored patients against survival models.

Table 2. Prediction performance results on the internal MUV dataset (Dynamic AUC and Concordance Index) comparing various deep models trained with LH and BCE loss, as well as a CoxPH model as baselines. Each model is compared with Deep LH using the estimate and confidence intervals (CI) obtained with bootstrapping. The best performance is denoted in bold. The asterisk * denotes a statistically significant difference compared to the proposed Deep LH.

Model	CCI (95% CI)	Dynamic AUC (95% CI)
CoxPH	0.78 [0.735 - 0.833] *	0.80 [0.749 - 0.849] *
Deep BCE 12m	0.58 [0.506 - 0.655] *	0.59 [0.512 - 0.662]*
Deep BCE 24m	0.67 [0.588 - 0.735] *	0.69 [0.614 - 0.761]*
Deep BCE 36m	0.70 [0.636 - 0.765] *	0.74 [0.683 - 0.800]*
Deep LH	0.77 [0.728 - 0.825]	0.79 [0.738 - 0.839]

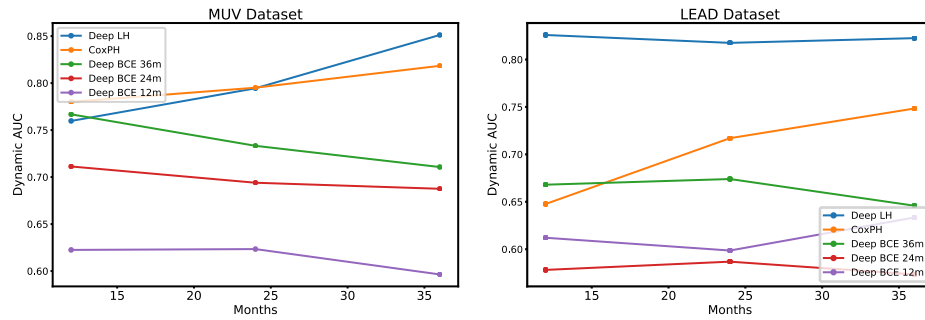


Fig. 6. Dynamic AUC curves from the cross-validation on the MUV and the LEAD dataset. AUC is evaluated for the 12, 24 and 36 months intervals. Five models are displayed: the deep survival model (Deep LH), three deep classification models (Deep BCE models) and the traditional survival model (CoxPH).

4.2. External validation

All the examined models were also evaluated on an external dataset (LEAD), both with CCI and dynamic AUC. The metrics were evaluated by combining the five models of the cross-validation, and their estimates and confidence intervals are displayed in Table 3 and in Fig. 6. The best performance was clearly obtained by the proposed Deep LH model in terms of both dynamic AUC with 0.82 (95% CI: 0.773 - 0.867), and CCI with 0.80 (95% CI: 0.745 - 0.846). All other models had a significant drop in performance compared to MUV dataset. The deep classification models struggled to generalize correctly to the external validation set. We also observed that the AUC performance was increasing for longer intervals for the survival models, while this was not the case for the binary classification models. Thus, the Deep LH model was able to generalize to an external dataset without decrease in the predictive performance.

Table 3. Prediction performance results (Dynamic AUC and Concordance Index) on the external validation set comparing various deep models trained with LH and BCE losses, as well as a CoxPH model as baselines. Each model is compared with Deep LH using the estimate and confidence intervals (CI) obtained with bootstrapping. The best performance is denoted in bold. The asterisk * denotes a statistically significant difference compared to the proposed Deep LH.

Model	CCI (95% CI)	Dynamic AUC (95% CI)
CoxPH	0.72 [0.642 - 0.797] *	0.72 [0.629 - 0.799] *
Deep BCE 12m	0.59 [0.549 - 0.638] *	0.62 [0.566 - 0.664] *
Deep BCE 24m	0.59 [0.534 - 0.635] *	0.58 [0.524 - 0.632] *
Deep BCE 36m	0.66 [0.601 - 0.722] *	0.66 [0.600 - 0.722] *
Deep LH	0.80 [0.745 - 0.846]	0.82 [0.773 - 0.867]

4.3. Risk group analysis

The Kaplan-Meier curves are displayed in Fig. 7 and the estimated progression rates in Table 4. Deep LH model, achieved the largest separation between low and high risk groups ($\Delta = 0.570$), while the intermediate risk groups were also correctly ordered. The CoxPh model had the best survival in low risk group, but more overlap with intermediate and high risk groups. On the other hand, Deep BCE models had a smaller separation, and the intermediate risk groups

overlapped with the low risk one. All the models achieved statistically significant separation between the low (bottom quartile) and high (top quartile) risk groups.

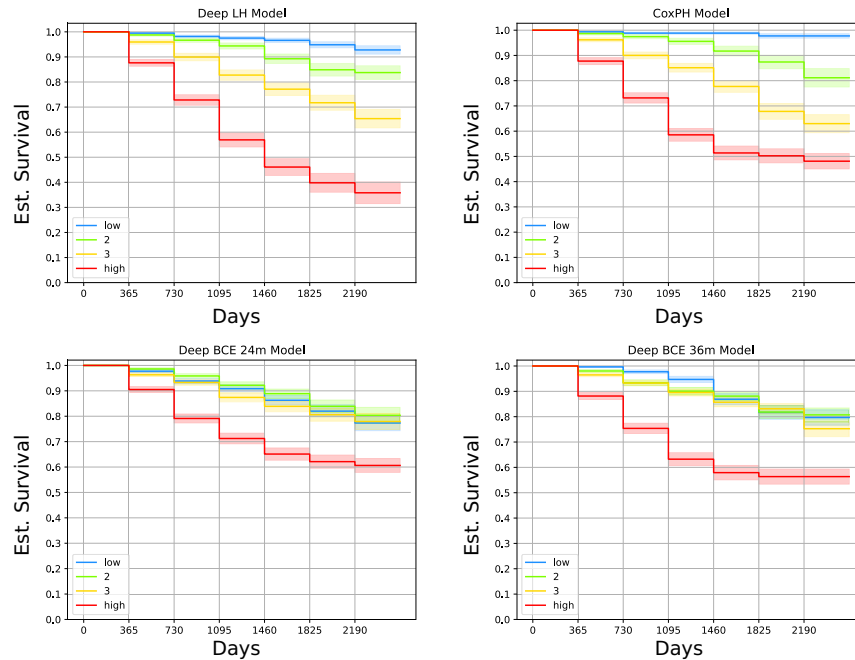


Fig. 7. Kaplan-Meier curves of the four risk groups predictions on the internal MUV dataset. We display two best deep classification models (Deep BCE 24m and Deep BCE 36m) as well as two survival models (CoxPH and Deep LH).

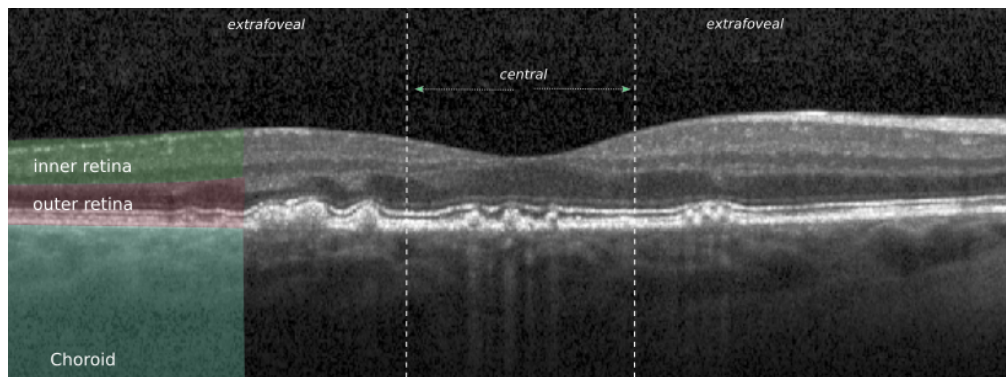


Fig. 8. Illustrations of regions to be occluded for the model interpretation experiments. The occluded region is filled with zeros and the prediction performance is computed on the MUV dataset with cross-validation.

4.4. Model interpretability with occlusion analysis

The occlusion sensitivity allows us to identify the regions of the retinal OCT that contribute the most evidence for the prediction at the cohort-level. The quantitative results are presented in Table 5. The largest drops were observed in two regions: the inner and outer retina (32.9 %

Table 4. Estimated progression rates in predicted risk groups on the internal MUV dataset. The progression rate as well as its variance are displayed. In addition, difference between the low risk and the high risk group are reported.

Model	Low risk (progression)	High risk (progression)	$S_{high} - S_{low}$
CoxPH	0.023 ± 0.0001	0.519 ± 0.0009	0.496 (p<0.001)
Deep BCE 12m	0.184 ± 0.0010	0.370 ± 0.0010	0.186 (p<0.001)
Deep BCE 24m	0.226 ± 0.0009	0.394 ± 0.0007	0.168 (p<0.001)
Deep BCE 36m	0.203 ± 0.0009	0.436 ± 0.0009	0.233 (p<0.001)
Deep LH	0.072 ± 0.0003	0.642 ± 0.0018	0.570 (p<0.001)

drop), both inside and outside the central mm region (19.0 % and 12.7%, respectively). The choroid region did not provide much evidence for the progression prediction. These experiments were coherent with the quantitative features used by the CoxPH model and previously reported prediction experiments [14]. Finally, at an individual-level, we show an example of CAM heatmaps for different prediction intervals (12, 24 and 36 months) in Fig. 9. We can observe a concentration of higher activations for 36 months hazard prediction in the central drusenoid area. In this case, the models correctly indentifies the time and the location of the conversion with lower activations in the 12 and 24 months hazards predictions.

Table 5. The average concordance index across the 5 folds on the MUV occluded dataset. The average concordance index was obtained by averaging across the test folds. Predictions were generated by Deep LH model.

Occlusion	Average C-Index	Relative
Fovea	0.64 ± 0.07	-19.0 %
Extrafovea	0.69 ± 0.05	-12.7 %
Choroid	0.74 ± 0.07	-6.3 %
Inner Retina	0.53 ± 0.10	-32.9 %
Outer Retina	0.53 ± 0.10	-32.9 %
No occlusion	0.79 ± 0.06	-

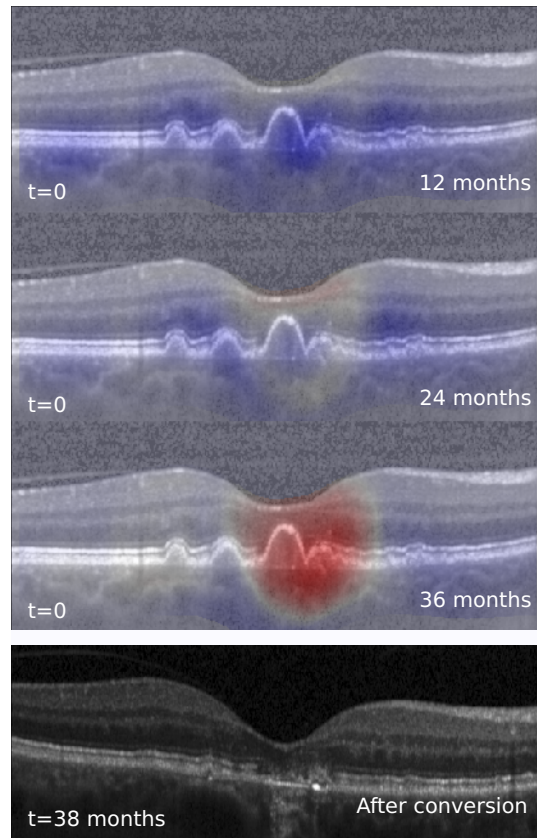


Fig. 9. An example of an eye that progresses in 38 months from baseline scan ($t = 0$) with CAM (class activation map). The first three rows are the activations for the baseline visit for 12, 24 and 36 months hazard prediction, the last row is the first available visit after the progression ($t = 38$ months). The hazards activation increase over time, and highlights the central drusenoid area where the future atrophy develops.

5. Discussion

The main challenge in predictive modeling from longitudinal data is that for many eyes, the progression event is not observed. In our datasets, they were either lost to follow-up or did not progress to the late stage of the disease within the respective study interval. This is further exacerbated by the generally slow advancement of AMD. To tackle these problems, we explored deep survival models based on Deep LH loss to predict progression from early/intermediate AMD to late atrophic AMD from retinal OCT. This task is of particular clinical relevance as the treatments based on complement inhibition are expected for late atrophic AMD [32,33], affecting more than 5 million people worldwide.

The proposed model has a deep learning component in the form of a CNN that enabled it to exploit the full extent the raw OCT imaging data offers and capture the predictive features hidden there. Unlike the majority of current deep learning approaches, we relied on a 3D CNN backbone, to account for the available volume-level labels. The LH training loss allowed to account for the information about the time to progression or patient censoring. To the best of our knowledge, this is the first deep learning model that was able to estimate the risk of progression to late atrophic AMD from raw 3D OCT volumes.

We compared our method to other standard deep learning solutions that consider the prediction task as a binary classification, i.e., whether an event has occurred or not within a given time interval. The proposed Deep LH models were able to outperform such baseline models both in terms of CCI and dynamic AUC, and on both the internal and more importantly, the external dataset. For Deep BCE models, the large amount of censored visits and the variability in times of progression (from 6 months to several years) limited the performance and indicated the benefits of using deep survival models. In addition, survival models have an advantage in training a single model as opposed to having to train a separate model for each desired prediction time interval.

We also compared our method to traditional Cox survival models based on a set of clinically relevant imaging biomarkers. Deep LH model performed on par with the CoxPH model on the internal dataset, but outperformed it on the important generalization to the external dataset. Furthermore, such biomarker-based models required the availability of several powerful image segmentation algorithms that are not necessary for the deep learning models. In addition, they ignore potentially relevant predictive imaging patterns not captured by the predefined set of quantitative features. Finally, because deep learning models require large datasets for training to learn to detect predictive patterns, as new datasets and the number of observed patients are expected to grow over time, the deep learning-based models are expected to keep improving correspondingly.

We validated the models on a large external dataset from a sham arm of a prospective trial to assess their generalization capability. In this challenging setting due to a population shift, our proposed approach remarkably preserved its performance, while the other methods showed decreased performance compared to the one on the internal dataset. In such a setting, the time-dependent LH loss contributed greatly over the standard BCE one. The standard deep classification models generalized poorly, especially the ones trained with short intervals. These results prove the limit of the binary classification approach for prediction, where the models failed to properly model the risk of progression. Despite the variability of observation periods and progression times in the training dataset, the LH survival loss allowed to properly capture the relevant longitudinal features.

We observed certain limitations that need further developments. Importantly, the proposed model had difficulty to predict the exact time of progression. This is partly because the discrete LH loss does not take into account the ordinal nature of the data. Similarly, the discretization of time requires setting a few hyperparameters, the length of the bins (time discretization) and the number of output neurons (defines the total duration covered by the network). We selected 6 intervals of 12 months, to get a good tradeoff between balance (limited number of progressors for each interval) and total duration of the MUV data (6 years). Improved losses or calibration methods may be required to get more precise estimations of individual time to progression. Furthermore, despite making use of large longitudinal datasets of natural progression of intermediate AMD, our training datasets remain quite small from a deep learning perspective, in particular, in the number of progressing eyes. Nevertheless, we chose to keep the two datasets separate to better assess the generalization capabilities of the models. In terms of the data, we identified three points that could be improved as part of future work: diversity of scanners, preprocessing step, and inclusion of wet AMD patients. We trained exclusively on scans from a Spectralis OCT, which limits the generalization of the deep learning models to other OCT devices, given their known sensitivity to image domain shift. Instead of performing ILM-flattening in the preprocessing, we could straighten according to Bruch's membrane, but its segmentation is challenging due to low intensity gradient under drusen and requires development of specialized algorithms. Finally, we excluded the very few eyes progressing to wet-AMD, as the conversion and the subsequent treatment with anti-VEGF drugs is expected to impact the risk of atrophy development. Thus, we expect that those should be modeled separately. However, the proposed survival model could be applied in the future to larger datasets of patients with wet-AMD cases.

Deep learning models are known to improve with larger datasets available for training, hence in the future, training on larger and more diverse studies will likely be needed to obtain performance that goes substantially beyond offered by currently clinically known predictive features. Thus, the ongoing large observational studies of patients with intermediate AMD, e.g., PINNACLE [34] (NCT04269304, ClinicalTrials.gov) and HONU (NCT05300724, ClinicalTrials.gov), will offer a high potential to obtain even more predictive deep learning systems of AMD progression. Finally, to further improve performance, specific CNN architectures that take into account the anisotropy of OCT scans, will be explored. Similarly, hybrid deep learning models that perform fusion of imaging and non-imaging features, e.g. patient demographic features with raw OCTs images, will be developed as part of future work.

In conclusion, the proposed deep survival model allowed training long-term prediction models directly from raw volumetric OCT scans. It accounted for the different times of progression and for the censoring of patients. Finally, it provided better prediction performance and more informative output than the commonly considered deep learning-based classification models.

Funding. Christian Doppler Research Association; Austrian Federal Ministry for Digital and Economic Affairs; National Foundation for Research, Technology and Development; Heidelberg Engineering.

Disclosures. Gregor Reiter: RetInSight (F). Robyn H. Guymer: Bayer (C), Novartis (C), Roche Genentech (C), Apellis (C). Ursula Schmidt-Erfurth: Genentech (F), Kodiak (F), Novartis (F), Apellis (F,C), RetinSight (F,P). Hrvoje Bogunović: Heidelberg Engineering (F), Apellis (F).

Data Availability. Data underlying the results presented in this paper are not publicly available at this time, but may be obtained from the authors upon reasonable request.

References

1. W. L. Wong, X. Su, X. Li, C. M. G. Cheung, R. Klein, C.-Y. Cheng, and T. Y. Wong, "Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis," *The Lancet. Glob. Heal.* **2**(2), e106–e116 (2014).
2. S. R. Sadda, R. Guymer, and F. G. Holz, *et al.*, "Consensus Definition for Atrophy Associated with Age-Related Macular Degeneration on OCT: Classification of Atrophy Report 3," *Ophthalmology* **125**(4), 537–548 (2018).
3. M. Nassisi, J. Lei, N. S. Abdelfattah, A. Karamat, S. Balasubramanian, W. Fan, A. Uji, K. M. Marion, K. Baker, X. Huang, E. Morgenthien, and S. R. Sadda, "OCT Risk Factors for Development of Late Age-Related Macular Degeneration in the Fellow Eyes of Patients Enrolled in the HARBOR Study," *Ophthalmology* **126**(12), 1667–1674 (2019).
4. J. Yim, R. Chopra, and T. Spitz, *et al.*, "Predicting conversion to wet age-related macular degeneration using deep learning," *Nat. Med.* **26**(6), 892–899 (2020).
5. R. H. Guymer, Z. Wu, and L. A. B. Hodgson, *et al.*, and Laser Intervention in Early Stages of Age-Related Macular Degeneration Study Group, "Subthreshold Nanosecond Laser Intervention in Age-Related Macular Degeneration: The LEAD Randomized Controlled Clinical Trial," *Ophthalmology* **126**(6), 829–838 (2019).
6. F. G. Schlanitz, B. Baumann, M. Kundi, S. Sacu, M. Baratsits, U. Scheschy, A. Shahlaee, T. J. Mittermüller, A. Montuoro, P. Roberts, M. Pircher, C. K. Hitzengerger, and U. Schmidt-Erfurth, "Drusen volume development over time and its relevance to the course of age-related macular degeneration," *Br. J. Ophthalmol.* **101**(2), 198–203 (2017).
7. E. M. Lad, K. Sleiman, and D. L. Banks, *et al.*, "Machine Learning OCT Predictors of Progression from Intermediate Age-Related Macular Degeneration to Geographic Atrophy and Vision Loss," *Ophthalmol. Sci.* **2**(2), 100160 (2022).
8. K. Sleiman, M. Veerappan, K. P. Winter, M. N. McCall, G. Yiu, S. Farsiu, E. Y. Chew, T. Clemons, and C. A. Toth, "Optical Coherence Tomography Predictors of Risk for Progression to Non-Neovascular Atrophic Age-Related Macular Degeneration," *Ophthalmology* **124**(12), 1764–1777 (2017).
9. U. Schmidt-Erfurth, S. M. Waldstein, S. Klmscha, A. Sadeghipour, X. Hu, B. S. Gerendas, A. Osborne, and H. Bogunović, "Prediction of Individual Disease Conversion in Early AMD Using Artificial Intelligence," *Invest. Ophthalmol. Vis. Sci.* **59**(8), 3199–3208 (2018).
10. D. B. Russakoff, A. Lamin, J. D. Oakley, A. M. Dubis, and S. Sivaprasad, "Deep Learning for Prediction of AMD Progression: A Pilot Study," *Invest. Ophthalmol. Vis. Sci.* **60**(2), 712–722 (2019).
11. I. Banerjee, L. de Sisternes, J. A. Hallak, T. Leng, A. Osborne, P. J. Rosenfeld, G. Gregori, M. Durbin, and D. Rubin, "Prediction of age-related macular degeneration disease using a sequential deep learning approach on longitudinal SD-OCT imaging biomarkers," *Sci. Rep.* **10**(1), 15434 (2020).
12. A. Bora, S. Balasubramanian, B. Babenko, S. Virmani, S. Venugopalan, A. Mitani, G. d. O. Marinho, J. Cuadros, P. Ruamviboonsuk, G. S. Corrado, L. Peng, D. R. Webster, A. V. Varadarajan, N. Hammel, Y. Liu, and P. Bavishi, "Predicting the risk of developing diabetic retinopathy using deep learning," *The Lancet Digit. Heal.* **3**(1), e10–e19 (2021).
13. D. R. Cox, "Regression Models and Life-Tables," *J. Royal Stat. Soc. Ser. B (Methodological)* **34**(2), 187–202 (1972).

14. Z. Wu, H. Bogunović, R. Asgari, U. Schmidt-Erfurth, and R. H. Guymer, "Predicting Progression of Age-Related Macular Degeneration Using OCT and Fundus Photography," *Ophthalmology Retina* **5**(2), 118–125 (2021).
15. M. Friedman, "Piecewise Exponential Models for Survival Data with Covariates," *Ann. Statist.* **10**(1), 101–113 (1982).
16. K. Liestøl, P. K. Andersen, and U. Andersen, "Survival analysis and neural nets," *Statist. Med.* **13**(12), 1189–1200 (1994).
17. D. Faraggi and R. Simon, "A neural network model for survival data," *Statist. Med.* **14**(1), 73–82 (1995).
18. H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Ann. Appl. Statist.* **2**(3), 841–860 (2008).
19. Y. Peng, T. D. Keenan, Q. Chen, E. Agrón, A. Allot, W. T. Wong, E. Y. Chew, and Z. Lu, "Predicting risk of late age-related macular degeneration using deep learning," *npj Digital Med.* **3**(1), 111 (2020).
20. J. Katzman, U. Shaham, J. Bates, A. Cloninger, T. Jiang, and Y. Kluger, "DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network," *BMC Med. Res. Methodol.* **18**(1), 24 (2018).
21. X. Zhu, J. Yao, and J. Huang, "Deep convolutional neural network for survival analysis with pathological images," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (IEEE, Shenzhen, China, 2016), pp. 544–547.
22. C. Lee, W. Zame, J. Yoon, and M. v. d. Schaar, "DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks," *Proc. AAAI Conf. on Artif. Intell.* **32**(1), 11842 (2018). Number: 1.
23. H. Kvamme and Ø. Borgan, "Continuous and discrete-time survival prediction with neural networks," *Lifetime Data Anal.* **27**(4), 710–736 (2021).
24. M. K. Garvin, M. D. Abramoff, X. Wu, S. R. Russell, T. L. Burns, and M. Sonka, "Automated 3-D Intraretinal Layer Segmentation of Macular Spectral-Domain Optical Coherence Tomography Images," *IEEE Trans. Med. Imaging* **28**(9), 1436–1447 (2009).
25. M. D. Abramoff, M. K. Garvin, and M. Sonka, "Retinal Imaging and Image Analysis," *IEEE Rev. Biomed. Eng.* **3**, 169–208 (2010). Conference Name: IEEE Reviews in Biomedical Engineering.
26. D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A Closer Look at Spatiotemporal Convolutions for Action Recognition," *arXiv*, arXiv:1711.11248 [cs] (2018). Number: arXiv:1711.11248.
27. I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *arXiv*, arXiv:1711.05101 [cs, math] (2019). Number: arXiv:1711.05101.
28. S. Pölsterl, "scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn," *J. Mach. Learn. Res.* **21**, 1–6 (2020).
29. H. Uno, T. Cai, L. Tian, and L. J. Wei, "Evaluating Prediction Rules for t-Year Survivors with Censored Regression Models," *J. Am. Stat. Assoc.* **102**(478), 527–537 (2007).
30. S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, (2010).
31. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," *arXiv*, arXiv:1512.04150 [cs] (2015). ArXiv: 1512.04150.
32. D. S. Liao, F. V. Grossi, and D. El Mehdi, *et al.*, "Complement C3 Inhibitor Pegcetacoplan for Geographic Atrophy Secondary to Age-Related Macular Degeneration: A Randomized Phase 2 Trial," *Ophthalmology* **127**(2), 186–195 (2020).
33. G. J. Jaffe, K. Westby, K. G. Csaky, J. Monés, J. A. Pearlman, S. S. Patel, B. C. Joondeph, J. Randolph, H. Masonson, and K. A. Rezaei, "C5 Inhibitor Avacincaptad Pegol for Geographic Atrophy Due to Age-Related Macular Degeneration: A Randomized Pivotal Phase 2/3 Trial," *Ophthalmology* **128**(4), 576–586 (2021).
34. J. Sutton, M. J. Menten, S. Riedl, H. Bogunović, O. Leingang, P. Anders, A. M. Hagag, S. Waldstein, A. Wilson, A. J. Cree, G. Traber, L. G. Fritsche, H. Scholl, D. Rueckert, U. Schmidt-Erfurth, S. Sivaprasad, T. Prevost, and A. Lotery, "Developing and validating a multivariable prediction model which predicts progression of intermediate to late age-related macular degeneration—the PINNACLE trial protocol," *Eye* **37**(6), 1275–1283 (2023).